The impact of segmentation on whole-lung functional MRI quantification: repeatability and reproducibility from multiple human observers and an artificial neural network.

Authors: Corin Willers[1], Grzegorz Bauman[2,3], Simon Andermatt[3], Francesco Santini[2,3], Robin Sandkühler[3], Kathryn Ramsey[1], Philippe C. Cattin[3], Oliver Bieri[2,3], Orso Pusterla*[2,3,4] and Philipp Latzin*[1]

* These authors contributed equally to the study and share last authorship.

Author Affiliations

[1]*Division of Pediatric Respiratory Medicine, Department of Pediatrics, Inselspital, Bern University Hospital, University of Bern, Switzerland*

[2]*Division of Radiological Physics, Department of Radiology, University of Basel Hospital, Basel, Switzerland*

[3]*Department of Biomedical Engineering, University of Basel, Basel, Switzerland*

[4]*Institute for Biomedical Engineering, University and ETH Zurich, Zurich, Switzerland*

Submitted to ***Magnetic Resonance in Medicine*** for possible publication as ***full paper***.

Corresponding author:

Philipp Latzin, MD, PhD

University Children's Hospital of Bern

Freiburgstrasse 8, 3010 Bern, Switzerland

E-mail: philipp.latzin@insel.ch

Phone: +41 31 632 9353

Paper details: Abstract 249 words. Paper Body 4613 words 6 Figures and 3 Tables. Supporting Information for review and online publication: 1 Supporting Figure.

Key Words: functional lung MRI, automated segmentation, neural networks, inter-reader reproducibility, pediatrics

ABSTRACT

Purpose: To investigate the repeatability and reproducibility of lung segmentation and their impact on the quantitative outcomes from functional pulmonary MRI. Additionally, to validate an artificial neural network (ANN) to accelerate whole-lung quantification.

Method: Ten healthy children and twenty-five children with cystic fibrosis underwent matrix pencil decomposition MRI (MP-MRI). Impaired relative fractional ventilation (R_{FV}) and relative perfusion (R_Q) from MP-MRI were compared using whole-lung segmentation performed by a physician at two time-points (A_{t1} and A_{t2}), by an MRI technician (B), and by an ANN (C). Repeatability and reproducibility were assess with dice similarity coefficient (DSC), paired t-test and Intraclass-correlation-coefficient (ICC).

Results: The repeatability within an observer (A_{t1} vs A_{t2}) resulted in a DSC of 0.94 ± 0.01 (mean ± SD), and an unsystematic difference of -0.01% for R_{FV} (p=0.92) and +0.1% for R_Q (p=0.21). The reproducibility between human observers (A_{t1} vs B) resulted in a DSC of 0.88 ± 0.02, and a systematic absolute difference of -0.81% (p<0.001) for R_{FV} and -0.38% (p=0.037) for R_Q. The reproducibility between human and the ANN (A_{t1} vs C) resulted in a DSC of 0.89 ± 0.03, and a systematic absolute difference of -0.36% for R_{FV} (p=0.017), and -0.35% for R_Q (p=0.002). The ICC was >0.98 for all variables and comparisons.

Conclusions: Despite high overall agreement, there were systematic difference in lung segmentation between observers. This needs to be considered for longitudinal studies and could be overcome by using an ANN, which performs as good as human observers and fully automatizes MP-MRI post-processing.

INTRODUCTION

Chronic pulmonary diseases remain one of the greatest public health challenges affecting both young and adult populations, severely reducing life quality and expectancy (1,2). Sensitive methods to diagnose and quantify early disease progression and treatment responses are crucial for effective respiratory medicine. In this regard, focus on the pediatric population is important as early respiratory events can impair lung development (3,4). Computed tomography (CT) represents the clinical gold standard for morphological lung assessment but the risk of cumulative radiation limits its application in longitudinal pediatric monitoring (5,6). Lung function tests such as the nitrogen multiple-breath washout technique (N2-MBW) may be sensitive to detect early ventilation inhomogeneity, but do not provide regional information which is needed to improve therapies and identify disease phenotypes (7-9).

Magnetic resonance imaging (MRI) of the lung is transitioning from the research setting to clinical application (10-12). One advantage of MRI is the ability to image functional deficits relating to ventilation and perfusion in the lung, which may be a more sensitive method to detect early disease than structural imaging (13). Commonly, lung ventilation imaging is performed through the inhalation of hyperpolarized or fluorinated gases, which necessitates specific equipment and highly trained personnel only available in specialized centers (14). Lung perfusion imaging is performed clinically by using intravenous gadolinium-based contrast agents, which can cause patient discomfort, increase complications for imaging, and subject patients to rare adverse events and health risks (15,16).

Recently, non-invasive proton-based (1H) functional imaging with Fourier decomposition (FD-MRI) and improved techniques such as matrix pencil decomposition (MP-MRI) and phase-resolved functional lung (PREFUL)-MRI allow simultaneous lung ventilation and perfusion assessment (17-19). These techniques require only free-breathing acquisitions and are therefore well tolerated and feasible in children. In patients with cystic fibrosis (CF), outcomes from the FD-MRI technique correlate with outcomes from spirometry and N2-MBW lung function tests (20-22). These associations were also reported in patients with chronic obstructive pulmonary disease (COPD), asthma, and bronchiectasis (23-25). Further, functional impairments observed on FD-MRI maps correlate with DCE and hyperpolarized helium-3 MRI (26,27). With the current results and the lack of specialized equipment, proton-based MRI has a high potential for clinical disease management.

For widespread clinical application, the quantification of outcomes from the MP-MRI need to be fully automatized. The generation of ventilation and perfusion weighted maps of the whole chest is automated with MP-MRI, and a radiologist can already assess these maps for whole-lung evaluation. However, quantitative assessment of ventilation and perfusion-defects

(currently given in percent) requires segmentation of the lung tissue. These segmentations are usually performed manually by a trained specialist, which is time-consuming and subjective. The impact of the observer on functional MRI outcomes is currently unknown.

For clinical application and use in longitudinal multicenter studies, it is essential to have robust, fast, quantitative, and reproducible post-processing of functional MRI outcomes which is observer-independent (28,29). Artificial neural networks (ANN) have made significant progress in discriminative tasks and displayed a high level of accuracy in organ segmentation (30,31). Notwithstanding the high performances of current deep learning models, segmentations performed by ANNs are still "observer-specific" (i.e. relative to the used and trained model) which can influence quantitative outcomes. We aimed to investigate the potential of ANNs to accelerate post-processing time for MP-MRI.

In this study, we investigate the intra-observer repeatability (variability across observations by the same observer) and inter-observer reproducibility (variability across different observers) of the segmentations and their subsequent impact on impaired relative ventilation and perfusion quantification of MP-MRI images. First, we determine the similarity of the segmentation masks performed by two human observers and an ANN. Further, we investigate the observer variability of MP-MRI quantitative outcomes and assess the applicability of the ANN for fully automatic evaluation.

METHODS

Study Design

This methodological study used existing MRI data from partly published measurements from a cross-sectional, single-center, observational study at the University Children's Hospital of Bern, Switzerland (20,21).

Study Population

In this study, data from 10 healthy controls and 25 children with CF were included. Eligibility criteria were the ability to perform pulmonary function tests and MRI, and no requirement for supplemental oxygen. Furthermore, for healthy subjects, no history of chronic lung disease and no acute respiratory infection in the 4 weeks prior to the investigations. Participants underwent N2-MBW, spirometry, body plethysmography, and MRI on the same day and in this order. Three patients underwent lung function testing and MR imaging twice resulting in a total of 38 examinations included in this study. From three patients two examinations were included, to increase the number of observations and thereby the precision of estimates. We considered

all the data as independent, as scans were performed at least one year apart. The study was approved by the Ethics Committee of Bern (EKNZ 2015-326 and KEK 2017-00279). Parents and participants gave written informed consent, if older than 14 years.

Lung Function Testing

N2-MBW was performed with an unmodified device (Exhalyzer D, Eco Medics AG, Duernten, Switzerland), and according to consensus guidelines (32). The primary outcome was the lung clearing index (LCI), which represents the ventilation inhomogeneity of the lung (calculated as the cumulative expired volume divided by functional residual capacity). Spirometry measurements (Jaeger MasterScreen, CareFusion, Hochberg, Germany) were performed after N2-MBW, according to ERS/ATS guidelines (33): The forced expiratory volume in 1s (FEV1) was used to describe the study population.

MRI Data Acquisition

Imaging was performed on a 1.5T whole-body MRI Scanner (MAGNETOM Aera; Siemens Healthineers, Erlangen, Germany) using a 12-channel thorax and a 24-channel spine receiver coil array. Children were awake and not sedated during the scans. Parents or caregivers were allowed to be with the child in the MR room during imaging.

State-of-the-art functional imaging for MP-MRI consisted of time-resolved two-dimensional (2D) coronal acquisitions (base-images) using an ultra-fast balanced steady-state free-precession pulse sequence (ufSSFP) during approximately 50 seconds of free-breathing (34). To cover the majority of lung, imaging was performed at 6-11 coronal slices positions. No contrast agent was used. Scan parameters for ufSSFP were as follow: field-of-view= 400×400 to 450×450 mm^2, matrix size = 128×128 (in-plane resolution = 3.1×3.1 to 3.5×3.5 mm^2), slice thickness = 12 mm, echo time / repetition time (TE/TR) = 0.67/1.46 ms, bandwidth = 2056 Hz/pixel, FA = 65°, GRAPPA factor 2, acquisition rate = 3.3 images/s, 160 coronal images per slice (= 48s scan time per slice) and predefined default shim settings (tune-up).

Functional Imaging with Matrix Pencil Decomposition

MP-MRI allows for robust calculation of ventilation and perfusion maps employing a matrix pencil decomposition and linearized least-square fitting for spectral analysis (18). It offers a fully automatic spectral analysis of the time-resolved data sets which improves the post-processing workflow. As compared to Fourier decomposition MRI, MP-MRI mitigates both the problem of time-series truncation and irregular breathing/cardiac frequencies.

Every acquired two-dimensional image time-series was registered to a fixed image chosen in the mid respiratory state (baseline image). Registration was performed using a specific algorithm which preserves ventilation and perfusion signal modulations but aligns automatically lung structures (e.g. vessels, chest cage, and airways) (35). Subsequently, the registered time-series was processed voxel-wise with the matrix pencil decomposition algorithm to calculate both perfusion-weighted and fractional ventilation maps. The first 10 images of every time series were excluded from the MP analysis since acquired in the transient state, and not in the pseudo-steady-state.

Quantification of Impaired Lung Functions

Segmentations of the lung parenchyma were performed on the baseline-images with the exclusion of the main pulmonary vessels. These segmentation masks were applied for pulmonary perfusion quantification, while for ventilation quantification the segmentation masks were refined as follows: the voxels in the lung perfusion maps representing the 95th percentile of signal intensities were identified as vessels and removed from the segmentation masks for ventilation quantification. Vessels are removed from the ventilation maps since they appear at low intensity and must be excluded for ventilation defect quantification. In order to extract distributions of fractional ventilation and perfusion, a threshold equal to 75% of the median value from each voxel distribution was used to quantify regions with impaired lung ventilation (R_{FV}) and perfusion (R_{Q}), as described before (26).

In this study, R_{FV} and R_{Q} were calculated in every subject as a lung-area-weighted average for the whole lung ($R_{FV,Lung}$ and $R_{Q,Lung}$). Figure 1 summarizes the steps required for MP-MRI and the quantitative evaluation of lung functions. The perfusion maps are normalized by the 97.5 percentile for graphical presentation as color maps.

Observers

Observer A (C.W.) is a physician with 1 year of experience in lung imaging. Next to segmentations drawn at time point one (A_{-t1}), observer A repeated the segmentations (A_{-t2}) after a minimum of 24 hours in a blinded random fashion to investigate the intra-observer repeatability. Observer B is an MR-technician with 5 years of experience in lung imaging. All human segmentations were done with a region-growing algorithm and manually refined in open-source software (MITK, version 1.1.0, DKFZ, Heidelberg, Germany) (36). Observer C represents the ANN.

Artificial Neural Network

An ANN (observer C) was trained to segment the lung parenchyma of baseline-images automatically (37). The recurrent neural network's main layers consist of multi-dimensional gated recurrent units (MD-GRU) for voxel-wise binary classification. Furthermore, on-the-fly data augmentation is applied during training to increase the network robustness, i.e. images and masks are both randomly and slightly scaled, rotated, skewed, distorted, noise is added, and image signal intensity is marginally variated. The MD-GRU neural network has already shown competitive accuracy for brain segmentation tasks, and specifically for lung segmentation, it previously reached a Dice similarity coefficient of 0.93 (37-39). The ANN can be found under https://github.com/zubata88/mdgru.

The artificial neural network was trained with baseline-images and lung segmentations of 51 patients examined multiple times in previous studies (totaling to 100 MR examinations, and 502 baseline images acquired in coronal orientation at several anterior-posterior positions.). Segmentations were originally drawn by an MR-scientist with 5 years of experience in lung imaging (O.P.). None of the subjects evaluated in this study was included in the ANN training data; the evaluation in this study represents thus a new and independent validation cohort for the network. The network training lasted 24 hours on a GPU (NVIDIA Quadro P6000, Nvidia Corp., Santa Clara, CA).

Data Analysis and Statistics

To evaluate the agreement between segmentation masks of two different observers ($S_{Observer1}$ and $S_{Observer2}$) or time points, we calculated the Dice similarity coefficient (DSC) as an overlay metric (40):

$$DSC=2*(S_{Observer1} \cap S_{Observer2}) / (S_{Observer1} + S_{Observer2}).$$

DSC ranges between [0,1], where zero indicates no overlap and 1 indicates exact overlap. DSC was calculated for single slices (DSC_{Slice}), and for the whole lung volume (DSC_{Lung}) of every patient. DSC was calculated between segmentations of observers A_{t1} and A_{t2}, A_{t1} and B, A_{t1} and C, B and C. For a qualitative assessment, an independent observer (O.P.) visually evaluated the automatically segmented data, determining which were not well performed.

Percentage differences between the indices resulting from different observers were computed as follows: $(R_{FV,Observer2} - R_{FV,Observer1}) / R_{FV,Observer1}*100$ and $(R_{Q,Observer2} - R_{Q,Observer1})/ R_{Q,Observer1}*100$. Absolute differences were assessed using a two-sided Wilcoxon signed-rank test, if normality assumption was not valid, otherwise paired t-test was used. The agreement and variability of R_{FV} and R_Q between observers were assessed graphically by the Bland–Altman method and by the coefficient of repeatability (CR) or reproducibility (RDC) (29,41). Reliability was assessed by calculating intra-class correlation (ICC) (42). The 95% limits of

agreement (LOA) between observers were calculated as the mean difference ± 1.96 SD. CR and RDC represent the least significant difference between to measurements. CR and RDC are estimated as 2.77 times the within-subject standard deviation through an analysis of variance, as described in (29). ICC estimates and their 95% confidence intervals were computed based on a mean-rating (k = 3), absolute-agreement, 2-way mixed-effects model. ICC was defined as excellent (>0.8), good (0.6-0.79), and moderate (0.4 – 0.59). The correlations between the functional defects (R_{FV} and R_Q) and the LCI were assessed with Pearson's correlation coefficient and linear regression models. For comparison of the non-nested models, we used adjusted R-squared and the Akaike Information Criterion (AIC) difference ($\Delta_i = AIC_i - AIC_{min}$) (43). We considered models with $\Delta_i < 2$ have no evidence against the *i*-th model. When $4 < \Delta_i < 7$ there is considerably less support for the *i*-th model. Models with $\Delta_i > 10$ were seen to have no support. Analyses were performed using Stata™ (StataCorp. 2015, Release 14. College Station, TX: StataCorp LP), Matlab (2017b, The MathWorks, Natick, MA) and GraphPad Prism (GraphPad Software Inc., La Jolla, CA). Two add on packages were used for Stata: BA-plot and Zanthro (44,45).

RESULTS

Study Population and Feasibility

Table 1 provides the general study population characteristics. The study population represents a broad cross-sectional spectrum of the pediatric age range from 6 to 18 years. All participants could perform spirometry, N2-MBW and MRI examinations. The R_{FV} for the healthy cohort included in this study was between 10.1% to 21.4% and 13.8% to 18.6% for R_Q (Observer At1) and for the CF cohort between 11.8% to 34.9% for R_{FV} and 14.4% to 34.5% for R_Q. The LCI ranged between 5.3 to 6.9 for the healthy, and between 7.4 to 18.2 for subjects with CF.

A total of 271 2D lung slices where segmented. Five exemplary selected segmentations performed by the three observers are presented in Figure 2, and additionally 15 randomly chosen segmentations are presented in the Supporting Information Figure S1 available online. All the segmentations are visually well performed and very similar, but the DSC vary between 0.82 and 0.96 (Figure 2), indicating that segmentations are subjective and that there is no absolute ground truth.

Qualitatively 94% of the segmentation performed by the ANN appeared well performed. In comparison, 6% had small imperfections which would require very minor manual refinement due to invasion of lung boundaries (e.g. chest, bowel [4%]), or the disease (e.g. atelectasis,

mucus [2%]) was partially not included in the lung mask. No manual refinement was performed in this study in order to evaluate fully automatic processing. Exemplary ANN segmentation flaws are presented in Figure 3.

Intra-observer Repeatability

Figure 4 shows boxplots of the DSC (DSC_{Slice} and DSC_{Lung}) for the intra-observer repeatability and the inter-observer reproducibility of the whole-cohort segmentations included in this study. The intra-observer repeatability for a segmentation mask has a very high similarity for both the single slices and the whole lung volume: DSC_{Slice} of 0.93 ± 0.04 (mean ± SD) and DSC_{Lung} of 0.94 ± 0.01. In Table 2 the t-test to evaluate the R_{FV} and R_Q biases is presented and in Figure 5 the Bland-Altman plots to evaluate the variability of R_{FV} and R_Q amongst the different observers. The high similarity in the intra-observer repeatability of segmentation results in an unsystematic absolute difference of -0.01% for R_{FV} (LOA, -1.2% to 1.1%) and +0.1% for R_Q (-0.8% to 1.0%). The CR for R_{FV} was 1.11% and 0.91% for R_Q. The ICC (95% CI) is very good for both ventilation with 0.996 (0.993 - 0.998) and perfusion 0.996 (0.993 – 0.998). The intra-observer repeatability for the ANN was perfect: DSC_{Slice} of 1.0 ± 0.0 and DSC_{Lung} of 1.0 ± 0.0.

Inter-observer Reproducibility

Between two human observers (A_{t1} and B), the segmentations show a good but not perfect similarity (Figure 4): DSC_{Slice} 0.87 ± 0.06 and DSC_{Lung} 0.88 ± 0.02. The difference in the segmentations of observer A_{t1} and B results in a small, but significant systematical bias of -0.81% in R_{FV} and -0.38% R_Q (Table 2). Bland-Altman analysis revealed a good agreement. As presented in Figure 5, between human observers, the 95% LOA was -2.5% to 0.9% for R_{FV} and -2.5% to 1.8% for R_Q. The RDC for R_{FV} was 2.33% and 2.20% for R_Q.

The DSC between the segmentations drawn by the human observer A_{t1} and the one computed by the ANN exhibits similar outcomes than the inter-observer reproducibility between two human observers: DSC_{Slice} 0.88 ± 0.05 and DSC_{Lung} 0.89 ± 0.03 (cf. Figure 4). Although having a good DSC, also the segmentations from the neural network introduce a small, systematic absolute difference of -0.36% in R_{FV} (LOA -2.1% to 1.4%) and -0.35% in R_Q (LOA -1.6% to 0.9%; cf. Table 2 and Figure 5). The RDC for the ANN against the human observer A was 1.86% for R_{FV} and 1.44% for R_Q. Comparing observer B to the ANN yielded similar results: DSC_{slice} 0.89 ± 0.06 and DSC_{Lung} 0.90 ± 0.03. There was a similar systematic difference for R_{FV} +0.45% (LOA -1.6% to 2.5%), and unsystematic difference for R_Q +0.03 (LOA -1.8% to 1.8%).

Notably, the human-ANN biases and the LOA are of similar extent as compared to the biases between two human observers. The ICC between the three observers shows a very good

agreement for both R_{FV} 0.984 (0.965 – 0.992) and R_Q 0.985 (0.975 – 0.992). In general, for all observers, the lowest DSC of the segmentation masks were observed in the very dorsal and the very ventral lung slices (see Figure 2 and Supporting Information Figure S1).

Impact on quantitative outcomes

In Figure 6 representative segmentations are shown for a subject with CF performed by the three observers and the resulting perfusion maps, perfusion defect masks, ventilation maps, and ventilation defect masks. Although the functional defect maps evaluated with segmentations outlined by different observers appear similar by eye (Figure 6), there is an impact on the quantification of impaired functions, namely varying from 27.8% to 28.5% in R_{FV} and from 25.2% to 26.4% in R_Q.

In the cohort of subjects presented in this study, the extent of R_{FV} and R_Q quantified by MP-MRI and by all three observers was strongly correlated to LCI. The Pearson correlation coefficients were r_{FV}= (0.772, 0.776, 0.775, 0.800) and r_Q=(0.785, 0.802, 0.773, 0.806), $p<0.0001$, for observers (A_{t1}, A_{t2}, B, C) respectively. Linear regression estimations of the LCI as a function of R_{FV} or R_Q and for the three observers are presented in Table 3. The intra-reader variation in the regression coefficients was higher for R_Q then for R_{FV}. The coefficient of determination R-squared (adjusted) was slightly higher for the artificial neural network. From the R-squared values, it is not possible to determine a relevant difference to give preference to an observer. The Akaike information criterion (AIC) and the AIC difference show considerably less support in the R_{FV} models from the human observers, and support the ANN the most. The AIC difference for R_Q varies from substantial support to considerably less support for the repeatability. Also for R_Q the ANN model had the most support.

DISCUSSION

Main Findings

The evaluation of intra-observer repeatability and inter-observer reproducibility are crucial to establish the stability and robustness of quantitative imaging outcomes for clinical application. For functional lung imaging, the segmentation is expected to have an impact on the quantitative outcomes. The human intra-observer repeatability for lung segmentation was excellent (DSC_{Lung}: 0.94) which resulted in a minimal and unsystematic influence on the imaging outcomes R_{FV} and R_Q. The ANN repeatability was perfect with a DSC_{Lung} of 1.0, and thereby did not show any bias.

The inter-observer reproducibility between two human observers show a good, but not perfect similarity (DSC: 0.88). This is essential to establish the general variability between observers, since medical image segmentation has the problem of lacking ground truth inherently and there is no knowledge towards a minimal significant difference of the DSC. The similarity between human and artificial neural network segmentations are on the same level of agreement (DSC: 0.89) as compared to human observers. The minimal difference in the segmentations introduces a small, but highly relevant bias between the quantitative outcomes R_{FV} and R_Q from different observers. This bias is of a similar extent between all observers. Moreover, the bias and limits of agreement for perfusion defects R_Q is lower compared to the ventilation defects R_{FV}, and due to the narrower distribution range of perfusion values as compared to fractional ventilation (not shown).

The overall agreement between the observers is good as shown by ICC. Independently from the observer, the correlation between LCI and both R_{FV} and R_Q was strong. The linear regression estimations of the LCI as a function of R_{FV} or R_Q for different observers were similar, but the Akaike information criterion supported the neural network the most.

Whether the inter-observer bias is relevant to the individual clinical case is yet to be determined. However, our findings demonstrate that the stability and continuity of the same observer for an interventional or longitudinal study are crucial. The longitudinal stability can be better controlled with an ANN, which represents a favorable advantage.

Comparison with previous studies

To the best of our knowledge, the intra-observer repeatability and inter-observer reproducibility of quantitative ventilation and perfusion lung defects as calculated from proton-based functional imaging have not yet been investigated. In previous work, we studied the reproducibility of MP-MRI measured 24-hour apart (21). The absolute differences of our previous 24-hour reproducibility study were R_Q=0.35% and R_{FV}=0.19% (segmentations performed by a single expert). In our current study, the repeatability differences caused by segmentations were lower (R_Q =0.1% ±0.46 and R_{FV}=0.01% ±0.57, mean ±SD), but interestingly the differences from human-human reproducibility were higher (R_Q = -0.38% ±1.07 and R_{FV}= -0.81% ±0.96). This corroborates the importance of having the same observer evaluating longitudinal data.

Our group has also proposed the application of a neural network for fully automated functional quantification. The lung is a frequently targeted organ for automated processing, but most applications focus on pathology or nodule detection on CT data, which is generally more widely used than MRI. Pulmonary MRI and the application of neural networks for automatic post-

processing remains still infrequent but rapidly growing (46). Recently Guo et. al. proposed an interesting automated processing for lung segmentation, but it was limited to proton-based ventilation MRI and still needs manually placed seeding points (47). Their approach reached a slightly better DSC (0.95 ± 0.01) for adult asthma patients. This value is of difficult comparison with our results due to the different MRI sequence used and different cohorts of patients. We used a bSSFP sequence optimized for lung imaging with a very short TR and focused on healthy and children with CF. Lately the original research of Tustison et. al. focused on the feasibility of proton-based MRI lung segmentations with an ANN, and their application to quantify pulmonary ventilation defects from hyperpolarized helium MRI in adult subjects (48). They reached a DSC of 0.94 ± 0.02, although they previously reported on even better performances (49). Our study differs since we investigated both ventilation and perfusion quantitative information without the application of contrast agents and without the use of hyperpolarized gases. Moreover, we focused on a pediatric population.

Strengths and Limitations

This validation cohort for the ANN represents a broad clinical spectrum of the pediatric CF population, which is an important target for early interventions. The data processed with the ANN represents real clinical data with no selection or preparation. The manual human segmentation is estimated to take about five to eight minutes per slice. With an average of seven slices per subject, a whole lung segmentation took about 35 to 57 minutes per patient. The ANN took only about three seconds to segment a single slice, totaling to less than 30 seconds for whole-lung segmentation. This gives an important advantage for the ANN.

The DSC provides similarity between segmentations, but it is sensitive to the size of the masks evaluated. In a small area, few false negative or false positive voxels have a strong influence on the coefficient. Therefore, it is challenging to achieve DSC >0.9 in the anterior or posterior regions/slices (smaller area) and generally young children with small lung volume (exemplary seen in Figure 2E). This explains the low values of DSC_{slice} and the rather wide range.

We acknowledge that our segmentations were not outlined by an experienced chest radiologist. Furthermore, the ANN segmentations might still have some minor flaws (see Figure 3). We account for this lack with our clinical workflow, as a chest radiologist is evaluating the functional maps overlaid onto the baseline images (as in Figure 6). Any flaws in the mask are easily identified and cause a return for refinement to the post-processing unit. Even with the best-performing automated model, validation and decisions are made by a physician.

Clinical relevance

Measures of repeatability and reproducibility are essential for the validation of imaging biomarkers. As imaging diagnostics include substantial post-processing, this technical aspect needs to be validated as thoroughly as the biological reproducibility. The reproducibility bias between the observers is of important consideration for studies with multiple investigations on one subject and needs to be quantified. In those studies where it is not possible to keep the observer (performing the segmentation) constant, it is crucial to quantify the reproducibility coefficient in order to estimate the impact of the inter-observer variability. In our study we could show, that a minimum change of 2.33% for R_{FV} and 2.20% for R_Q would be above the inter-observer variability.

Although current results of functional lung MRI in general and MP-MRI specifically are very promising, one clear drawback for clinical routine use is the time-consuming post-processing procedure. To incorporate imaging biomarkers into clinical decision making it needs to be quickly available to the treating physician. The here demonstrated ANN can work time and resource-efficient, delivering results within minutes to the clinicians and patients. Although only CF patients were investigated, our findings apply to other lung diseases such as COPD, Asthma or lung cancer.

Methodological considerations & next steps

In this study, we explored the influence of the segmentation masks on the quantitative outcome. Repeatability for a human observer results in limits of agreements of ±1.2% for R_{FV} and ±0.9% for R_Q (Figure 5) which correspond to the uncertainty caused only by the segmentations variability. The repeated use of the ANN produced exactly the same segmentation results, therefore we expect that in a scan-rescan experiment any change in R_{FV} or R_Q would be caused by other experimental uncertainties such as patient and coil positioning, MR hardware settings (shimming, frequency adjustments), and physiology (breathing and cardiac pulsation). These experimental uncertainties of the MP-MRI imaging method must yet be determined by scan-rescan reproducibility studies and taken into account especially for longitudinal patients evaluation. Due to the high repeatability of the ANN segmentations, we expect that in longitudinal examinations the ANN mitigates segmentation uncertainties and improves overall repeatability as compared to segmentations performed by humans.

Recurrent neural networks are able to meet or even surpass the state-of-the-art results of feed-forward convolutional neural networks (CNN, e.g. U-net) (38,39). However, it might be interesting to test a feed-forward convolutional neural network for our specific segmentation task or to improve our current network with a more elaborate on-the-fly data augmentation technique (48). There is no inherent ground truth available for organ segmentation. Using

segmentations of several observers or consensus segmentations as target for ANN training might reduce the dependence of the ANN output on different segmentation “styles" of different observers and generalize better. The ideal number of human observers needed to create training data is unclear and depends upon several factors, such as underlying organ, training process of the human observers and resulting variability of segmentation between observers. Moreover, currently every research facility is gathering its own data and developing its own models for this task. It is tempting to propose joining resources for large-scale benefits, as it has been done for other anatomical locations (50).

CONCLUSION

Segmentations of the lung are highly repeatable without any bias. Human observers and an ANN reproduce the task well, but observers are not interchangeable since reproducibility introduces a systematic bias in the resulting quantifications. This drawback can be eliminated by the proposed ANN, which processes the data time and resource-efficient with similar accuracy to human observers, strengthening its routine use. Since the post-processing is now fully automatized for MP-MRI, we see a potential for broad clinical application in several lung diseases.

Acknowledgements: This work was kindly supported by the Swiss National Science Foundation (grant number: 182719 and 168173). Orso Pusterla acknowledges the financial support from the Strategic Focus Area initiative “Personalized Health and Related Technologies (PHRT, grant #2018-223)” of the ETH Domain, Switzerland. The authors would like to thank all children and their families for participation in the study. The authors would like to express their thankfulness especially to Mrs. Wirz, Mrs. Lüscher, Mrs. Krattinger, Mrs. Haas, Mrs. Beutler-Minth, Dr. Ith, to all the medical-technical assistants from the radiology department for their patient care and support in measurements.

REFERENCES

1. Bousquet J, Kaltaev N. Global surveillance, prevention and control of chronic respiratory diseases : a comprehensive approach / edited by Jean Bousquet and Nikolai Khaltaev. In. Geneva: World Health Organization; 2007.
2. Ley-Zaporozhan J, Ley S, Kauczor H-U. Morphological and functional imaging in COPD with CT and MRI: present and future. *European radiology.* 2008;18(3):510-521.
3. Wielpütz M, Kauczor H-U. MRI of the lung: state of the art. *Diagn Interv Radiol.* 2012;18(4):344-353.
4. Bui DS, Lodge CJ, Burgess JA, et al. Childhood predictors of lung function trajectories and future COPD risk: a prospective cohort study from the first to the sixth decade of life. *The Lancet Respiratory Medicine.* 2018;6(7):535-544.
5. Pearce MS, Salotti JA, Little MP, et al. Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. *Lancet.* 2012;380(9840):499-505.
6. Kuo W, Ciet P, Tiddens HAWM, Zhang W, Guillerman RP, van Straten M. Monitoring cystic fibrosis lung disease by computed tomography. Radiation risk in perspective. *American journal of respiratory and critical care medicine.* 2014;189(11):1328-1336.
7. Kieninger E, Singer F, Fuchs O, et al. Long-term course of lung clearance index between infancy and school-age in cystic fibrosis subjects. *Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society.* 2011;10(6):487-490.
8. Gustafsson PM, De Jong PA, Tiddens HA, Lindblad A. Multiple-breath inert gas washout and spirometry versus structural lung disease in cystic fibrosis. *Thorax.* 2008;63(2):129-134.
9. Verbanck S, Schuermans D, Paiva M, Meysman M, Vincken W. Small airway function improvement after smoking cessation in smokers without airway obstruction. *American journal of respiratory and critical care medicine.* 2006;174(8):853-857.
10. Wielputz MO, Mall MA. Imaging modalities in cystic fibrosis emerging role of MRI. *Pulmonary Medicine.* 2015.
11. Mall MA, Stahl M, Graeber SY, Sommerburg O, Kauczor HU, Wielputz MO. Early detection and sensitive monitoring of CF lung disease: Prospects of improved and safer imaging. *Pediatric pulmonology.* 2016;51(S44):S49-S60.
12. Roach DJ, Cremillieux Y, Fleck RJ, et al. Ultrashort Echo-Time Magnetic Resonance Imaging Is a Sensitive Method for the Evaluation of Early Cystic Fibrosis Lung Disease. *Annals of the American Thoracic Society.* 2016;13(11):1923-1931.
13. Wielpütz MO, Puderbach M, Kopp-Schneider A, et al. Magnetic Resonance Imaging Detects Changes in Structure and Perfusion, and Response to Therapy in Early Cystic Fibrosis Lung Disease. *American journal of respiratory and critical care medicine.* 2014;189(8):956-965.
14. Smith LJ, Collier GJ, Marshall H, et al. Patterns of regional lung physiology in cystic fibrosis using ventilation magnetic resonance imaging and multiple-breath washout. *The European respiratory journal.* 2018;52(5):1800821.
15. Granata V, Cascella M, Fusco R, et al. Immediate Adverse Reactions to Gadolinium-Based MR Contrast Media: A Retrospective Analysis on 10,608 Examinations. *Biomed Res Int.* 2016;2016:3918292-3918292.
16. Stahl M, Wielputz MO, Graeber SY, et al. Comparison of Lung Clearance Index and Magnetic Resonance Imaging for Assessment of Lung Disease in Children with Cystic Fibrosis. *American journal of respiratory and critical care medicine.* 2017;195(3):349-359.

17. Bauman G, Puderbach M, Deimling M, et al. Non-contrast-enhanced perfusion and ventilation assessment of the human lung by means of fourier decomposition in proton MRI. *Magnetic resonance in medicine.* 2009;62(3):656-664.
18. Bauman G, Bieri O. Matrix pencil decomposition of time-resolved proton MRI for robust and improved assessment of pulmonary ventilation and perfusion. *Magnetic resonance in medicine.* 2017;77(1):336-342.
19. Voskrebenzev A, Gutberlet M, Klimes F, et al. Feasibility of quantitative regional ventilation and perfusion mapping with phase-resolved functional lung (PREFUL) MRI in healthy volunteers and COPD, CTEPH, and CF patients. *Magnetic resonance in medicine.* 2018;79(4):2306-2314.
20. Nyilas S, Bauman G, Sommer G, et al. Novel magnetic resonance technique for functional imaging of cystic fibrosis lung disease. *The European respiratory journal.* 2017;50(6).
21. Nyilas S, Bauman G, Pusterla O, et al. Ventilation and perfusion assessed by functional MRI in children with CF: reproducibility in comparison to lung function. *Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society.* 2018.
22. Nyilas S, Bauman G, Pusterla O, et al. Structural and Functional Lung Impairment in PCD: Assessment with MRI and Multiple Breath Washout in Comparison to Spirometry. *Annals of the American Thoracic Society.* 2018.
23. Capaldi DP, Sheikh K, Guo F, et al. Free-breathing pulmonary 1H and Hyperpolarized 3He MRI: comparison in COPD and bronchiectasis. *Academic radiology.* 2015;22(3):320-329.
24. Capaldi DPI, Eddy RL, Svenningsen S, et al. Free-breathing Pulmonary MR Imaging to Quantify Regional Ventilation. *Radiology.* 2018;287(2):693-704.
25. Kaireit TF, Voskrebenzev A, Gutberlet M, et al. Comparison of quantitative regional perfusion-weighted phase resolved functional lung (PREFUL) MRI with dynamic gadolinium-enhanced regional pulmonary perfusion MRI in COPD patients. *Journal of magnetic resonance imaging : JMRI.* 2018.
26. Bauman G, Puderbach M, Heimann T, et al. Validation of Fourier decomposition MRI with dynamic contrast-enhanced MRI using visual and automated scoring of pulmonary perfusion in young cystic fibrosis patients. *European journal of radiology.* 2013;82(12):2371-2377.
27. Bauman G, Scholz A, Rivoire J, et al. Lung ventilation- and perfusion-weighted Fourier decomposition magnetic resonance imaging: in vivo validation with hyperpolarized 3He and dynamic contrast-enhanced MRI. *Magnetic resonance in medicine.* 2013;69(1):229-237.
28. Abramson RG, Burton KR, Yu JP, et al. Methods and challenges in quantitative imaging biomarker development. *Academic radiology.* 2015;22(1):25-32.
29. Raunig DL, McShane LM, Pennello G, et al. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. *Statistical methods in medical research.* 2015;24(1):27-67.
30. Christe A, Peters AA, Drakopoulos D, et al. Computer-Aided Diagnosis of Pulmonary Fibrosis Using Deep Learning and CT Images. *Investigative radiology.* 2019;54(10):627-632.
31. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436-444.
32. Robinson PD, Latzin P, Verbanck S, et al. Consensus statement for inert gas washout measurement using multiple- and single- breath tests. *The European respiratory journal.* 2013;41(3):507-522.
33. Miller MR, Hankinson J, Brusasco V, et al. Standardisation of spirometry. *The European respiratory journal.* 2005;26(2):319-338.
34. Bauman G, Pusterla O, Bieri O. Ultra-fast Steady-State Free Precession Pulse Sequence for Fourier Decomposition Pulmonary MRI. *Magnetic resonance in medicine.* 2016;75(4):1647-1653.

35. Sandkühler R, Jud C, Pezold S, Cattin PC. Adaptive Graph Diffusion Regularisation for Discontinuity Preserving Image Registration. Paper presented at: 8th International Workshop on Biomedical Image Registration (WBIR)2018. doi:10.1007/978-3-319-92258-4_3
36. Maleike D, Nolden M, Meinzer HP, Wolf I. Interactive segmentation framework of the Medical Imaging Interaction Toolkit. *Computer methods and programs in biomedicine.* 2009;96(1):72-83.
37. Pusterla O, Andermatt S, Bauman G, et al. Deep Learning Lung Segmentation in Paediatric Patients. Paper presented at: Proceedings of the 26th Annual Meeting of ISMRM; 2018, 2018; Paris, France.
38. Andermatt S, Pezold S, Cattin PC. Automated Segmentation of Multiple Sclerosis Lesions Using Multi-dimensional Gated Recurrent Units. Paper presented at: MICCAI Workshop Deep Learning in Medical Image Analysis2018; Cham. doi:https://doi.org/10.1007/978-3-319-75238-9_3
39. Andermatt S. PS, Cattin P. Multi-dimensional Gated Recurrent Units for the Segmentation of Biomedical 3D-Data BT. *Deep Learning and Data Labeling for Medical Applications.* 2016(10008):142-151.
40. Dice LR. Measures of the Amount of Ecologic Association Between Species. *Ecology.* 1945;26(3):297-302.
41. Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. *Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology.* 2003;22(1):85-93.
42. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med.* 2016;15(2):155-163.
43. Burnham KP, Anderson DR. A practical information-theoretic approach. *Model selection and multimodel inference, 2nd ed Springer, New York.* 2002.
44. *BAPLOT: Stata module to produce Bland-Altman plots* [computer program]. 2014.
45. Vidmar SI, Cole TJ, Pan H. Standardizing anthropometric measures in children and adolescents with functions for egen: Update. *Stata Journal.* 2013;13(2):366-378.
46. Lenchik L, Heacock L, Weaver AA, et al. Automated Segmentation of Tissues Using CT and MRI: A Systematic Review. *Academic radiology.* 2019.
47. Guo F, Capaldi DPI, McCormack DG, Fenster A, Parraga G. A framework for Fourier-decomposition free-breathing pulmonary (1) H MRI ventilation measurements. *Magnetic resonance in medicine.* 2018.
48. Tustison NJ, Avants BB, Lin Z, et al. Convolutional Neural Networks with Template-Based Data Augmentation for Functional Lung Image Quantification. *Academic radiology.* 2019;26(3):412-423.
49. Tustison NJ, Qing K, Wang C, Altes TA, Mugler JP, 3rd. Atlas-based estimation of lung and lobar anatomy in proton MRI. *Magnetic resonance in medicine.* 2016;76(1):315-320.
50. Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging.* 2015;34(10):1993-2024.

TABLES

	Healthy (n = 10)	Cystic Fibrosis (n = 25)
Age, years	11.1 (± 4.0)	13.5 (± 3.6)
Range	5.7 to 17.2	6.1 to 18.9
Males in % (N)	60% (n=6)	40% (n=10)
Weight (kg)	39.6 (± 17.1)	43.8 (± 13.7)
Weight [z-score]	0.2 (± 0.4)	-0.4 (± 0.8)
Height (meter)	1.4 (± 0.2)	1.5 (± 0.2)
Height [z-score]	-0.1 (± 1.0)	-0.4 (± 1.1)
BMI [z-score]	0.4 (± 0.6)	-0.2 (± 0.7)
FEV_1 [z-score]	0.4 (± 0.8)	-1.7 (± 1.5)
LCI	6.1 (± 0.5)	11.7 (± 2.9)
R_{FV} (from Observer A_{t1})	16.1 (± 4.2)	24.1 (± 6.3)
Range	10.1 to 21.4	11.8 to 34.9
R_Q (from Observer A_{t1})	16.9 (± 1.5)	23.4 (± 5.4)
Range	13.7 to 18.6	14.4 to 34.5

Table 1. General population statistics of the study cohort in healthy and cystic fibrosis participants for demographics and pulmonary function tests. Note: Each value is given as mean and standard deviation (SD), if not stated otherwise. Abbreviations: BMI, body mass index; FEV_1, forced expiratory volume in 1 second; LCI, Lung clearance index; R_{FV}, impaired relative fractional ventilation; R_Q, impaired relative perfusion z-scores according to WHO data from Stata 15 (45).

	Observers	Defect Percentage (Mean)	Defect Percentage (Mean)	Absolute difference (± SD)	95% Confidence Interval	P-value
Ventilation (R_{FV})	A_{t1} - A_{t2}	21.97	21.98	-0.01 (± 0.57)	-0.2 to 0.18	0.921
	A_{t1} - B	21.97	22.78	-0.81 (± 0.96)	-1.10 to -0.52	<0.001
	A_{t1} - C	21.97	22.33	-0.36 (± 0.89)	-0.65 to -0.7	0.017
	B - C	22.83	22.33	0.45 (± 1.17)	0.11 to 0.79	0.011
Perfusion (R_Q)	A_{t1} - A_{t2}	21.69	21.60	0.096 (± 0.46)	-0.06 to 0.25	0.211
	A_{t1} - B	21.69	22.07	-0.38 (± 1.07)	-0.73 to -0.02	0.037
	A_{t1} - C	21.69	22.04	-0.35 (± 0.65)	-0.56 to -0.14	0.002
	B - C	22.07	22.04	0.03 (± 0.96)	-0.29 to 0.34	0.87

Table 2. Paired T-Test between observers for R_{FV} and R_Q. Intra-observer repeatability shows a small, non-significant difference for both R_{FV} and R_Q. Human – human difference is of similar extent as human-ANN. Note: Defect size is given as percentage of whole lung. Abbreviations: R_{FV}, impaired relative fractional ventilation; R_Q, impaired relative perfusion.

Independent variable	Model	Adjusted R-squared	AIC	AIC difference ($\Delta_i = AIC_i - AIC_{min}$)
A_{t1}	LCI = (1.44 ±1.32) + (0.39 ±0.06)*R_{FV}	0.58	155.7	4
A_{t2}	LCI = (1.32 ±1.32) + (0.40 ±0.06)*R_{FV}	0.59	155.3	3.6
B	LCI = (0.53 ±1.43) + (0.42 ±0.06)*R_{FV}	0.59	155.3	3.6
C	LCI = (0.53 ±1.32) + (0.43 ±0.06)*R_{FV}	0.63	151.7	0
A_{t1}	LCI = (-0.65 ±1.55) + (0.49 ±0.07)*R_Q	0.60	153.9	3
A_{t2}	LCI = (-1.01 ±1.51) + (0.51 ±0.07)*R_Q	0.63	151.5	0.6
B	LCI = (-0.89 ±1.64) + (0.50 ±0.07)*R_Q	0.58	155.6	4.7
C	LCI = (-0.97 ±1.48) + (0.50 ±0.06)*R_Q	0.64	150.9	0

Table 3. Linear regression analyses: effect of R_{FV} and R_Q from different observers on correlation to LCI. Model equations for different observers as independent and LCI as dependent variable, including adjusted R-squared, AIC and AIC difference. The slope for the models is very similar for the observers. The adjusted R-squared and the AIC support the ANN segmentations at most. Abbreviations: R_{FV}, impaired relative fractional ventilation; R_Q, impaired relative perfusion; LCI, Lung clearance index; AIC, Akaike Information Criterion; ANN, artificial neural network.

FIGURES

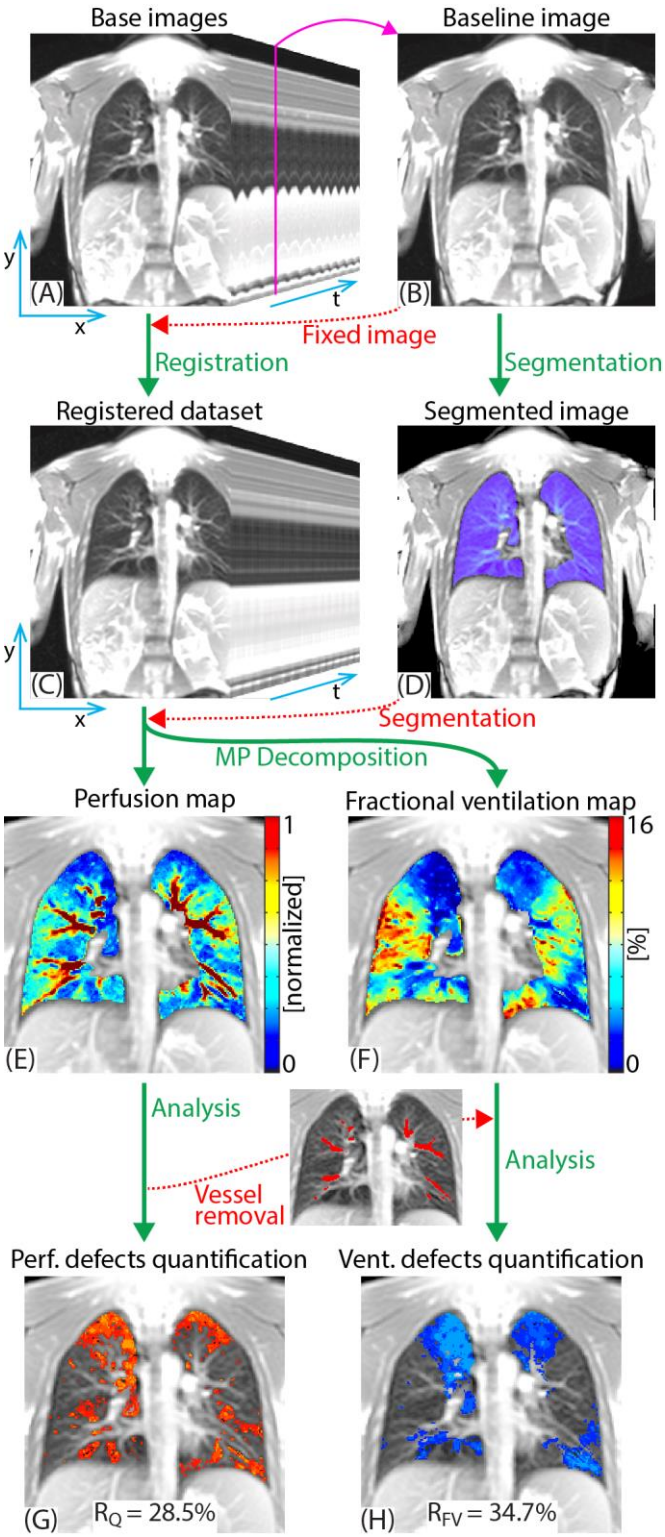


Figure 1. Graphical overview of the workflow to quantify R_Q and R_{FV}. Time resolved base images (A) are registered (C) on a fixed baseline image in the middle of the respiratory state (B). The baseline image (B) is used to segment the lung (D). The registered time-series (C) is processed with the matrix pencil decomposition algorithm to calculate perfusion and ventilation

maps (E,F), and the lung segmentation (D) is applied to mask the lung. Histogram distribution analysis allows to visualize and quantify perfusion and ventilation defects (G,H). To note, the high intensity appearing lung vessels on perfusion maps (E) are removed from the lung segmentation to quantify the ventilation defects (workflow from F to H, cf. “Quantification of Impaired Lung Functions” in the “Methods” section for more information). The segmentation in (D) is outlined by three different observers to investigate the impact of different observers on R_Q and R_{FV}. In this subject with CF (17 years old, male, FEV1 z-score= -4.1, LCI = 14.4) impaired lung perfusion and ventilation are 28.5% and 34.7% respectively (G and H). Abbreviations: FEV_1, forced expiratory volume in 1 second; LCI, Lung clearance index; R_{FV}, impaired relative fractional ventilation; R_Q, impaired relative perfusion.

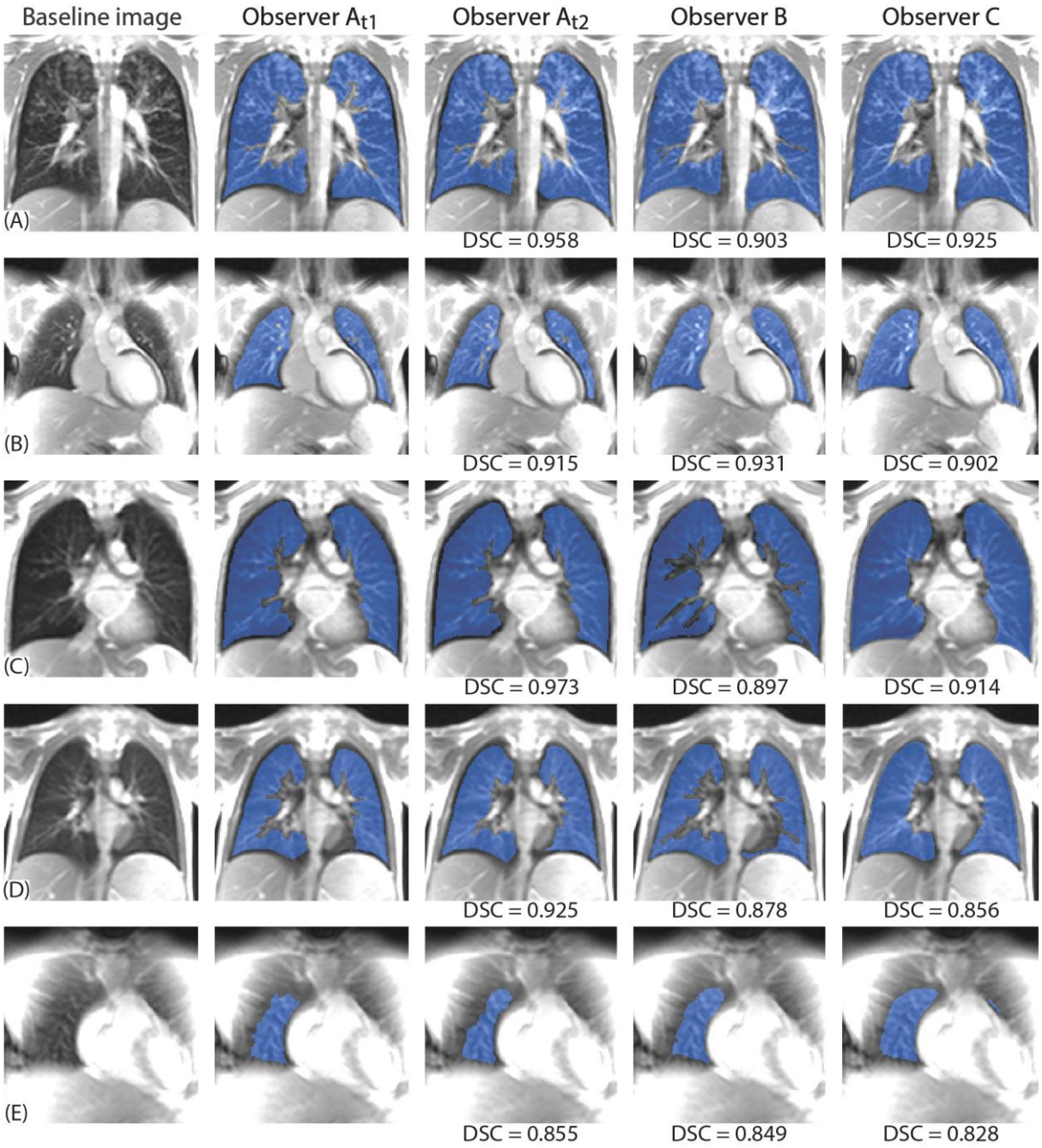


Figure 2. Segmentations in 5 subjects (A-E) outlined by the 3 different observers including A_{t2}. Segmentations masks are overlaid in blue onto greyscale baseline-image. Segmentations amongst different observers are very similar and all appear correct, emphasizing there is no ground truth for this task. To note observer A seems to be more conservative in including pixels next to the lung borders and diaphragm, as compared to observers B and C. Row (A) and (B) represent examples mostly above the DSC average. Row (C) and (D) represent examples around the DSC average. Row (E) represents an example below the DSC average: Although by eyeballing quite similar between observers, theses segmentations yield a low DSC due to

the small mask size (N of mask pixels ca. 900 in E as compared to 10’000 for A). The DSC is calculated using observer A_{t1} as reference. Abbreviations: DSC, dice similarity coefficient.

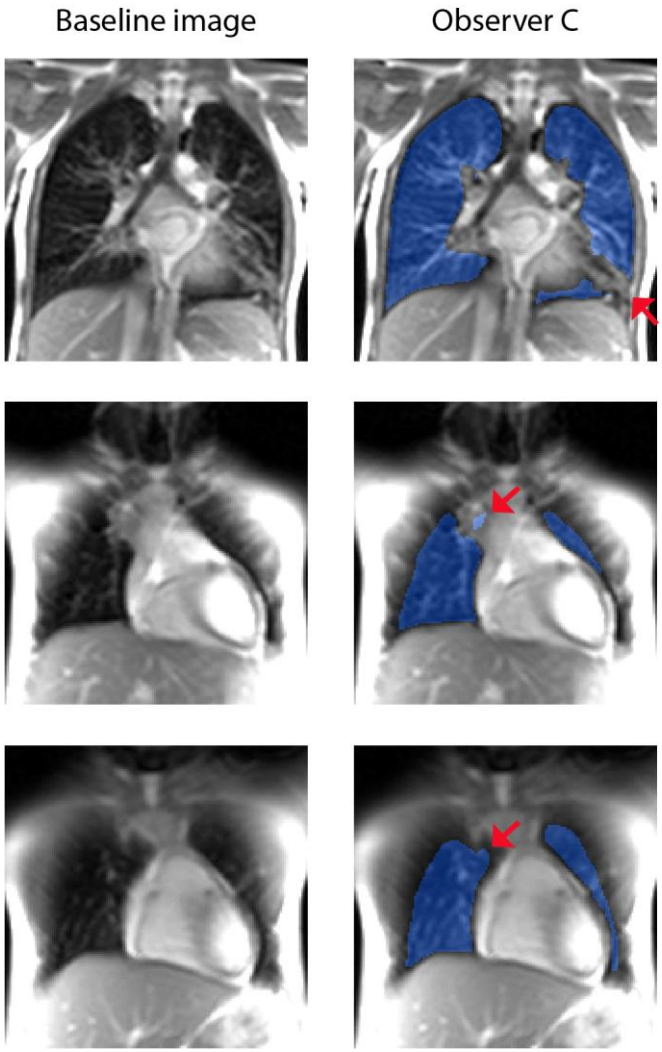


Figure 3. Exemplary segmentation flaws of the ANN. The arrows indicate segmentation flaws of boundaries caused by lung atelectasis not included in the segmentation (firs row), and inclusion of non-lung tissue (middle row) and partial volume (last row). Abbreviations: ANN, artificial neural network.

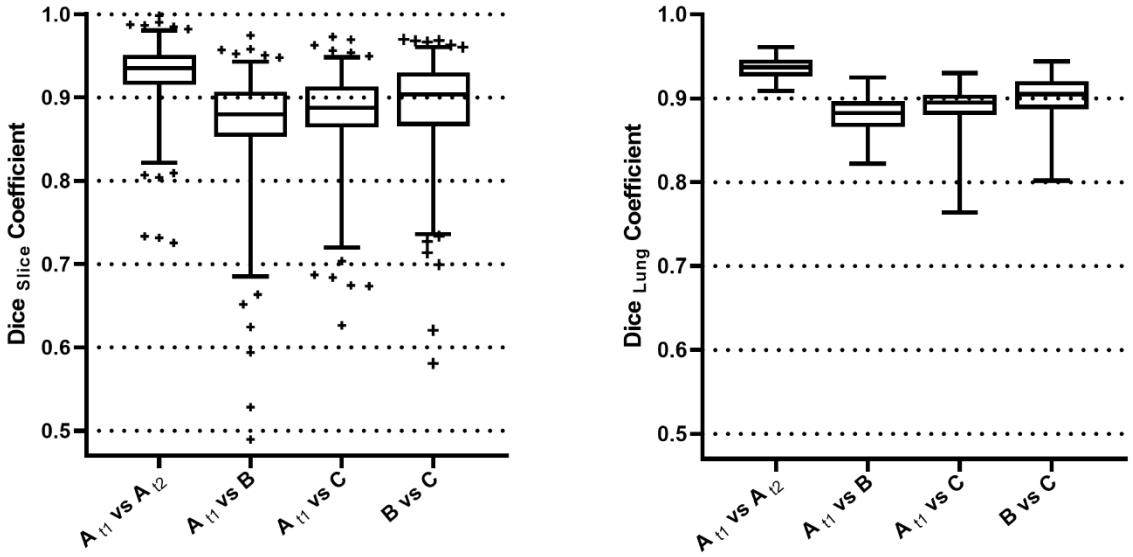


Figure 4. Boxplot, Whiskers: 2.5 to 97.5 Percentile. DSC for four distinct Observers (A_{t1}, A_{t2}, B and C): On the left DSC for 271 single coronal slices, and on the right DSC for 38 lung volumes. The mean DSC between human and the ANN segmentations (A vs. C, and B vs C) is of similar extent as compared to the DSC between human observers (A vs B). Abbreviations: DSC, dice similarity coefficient; ANN, artificial neural network.

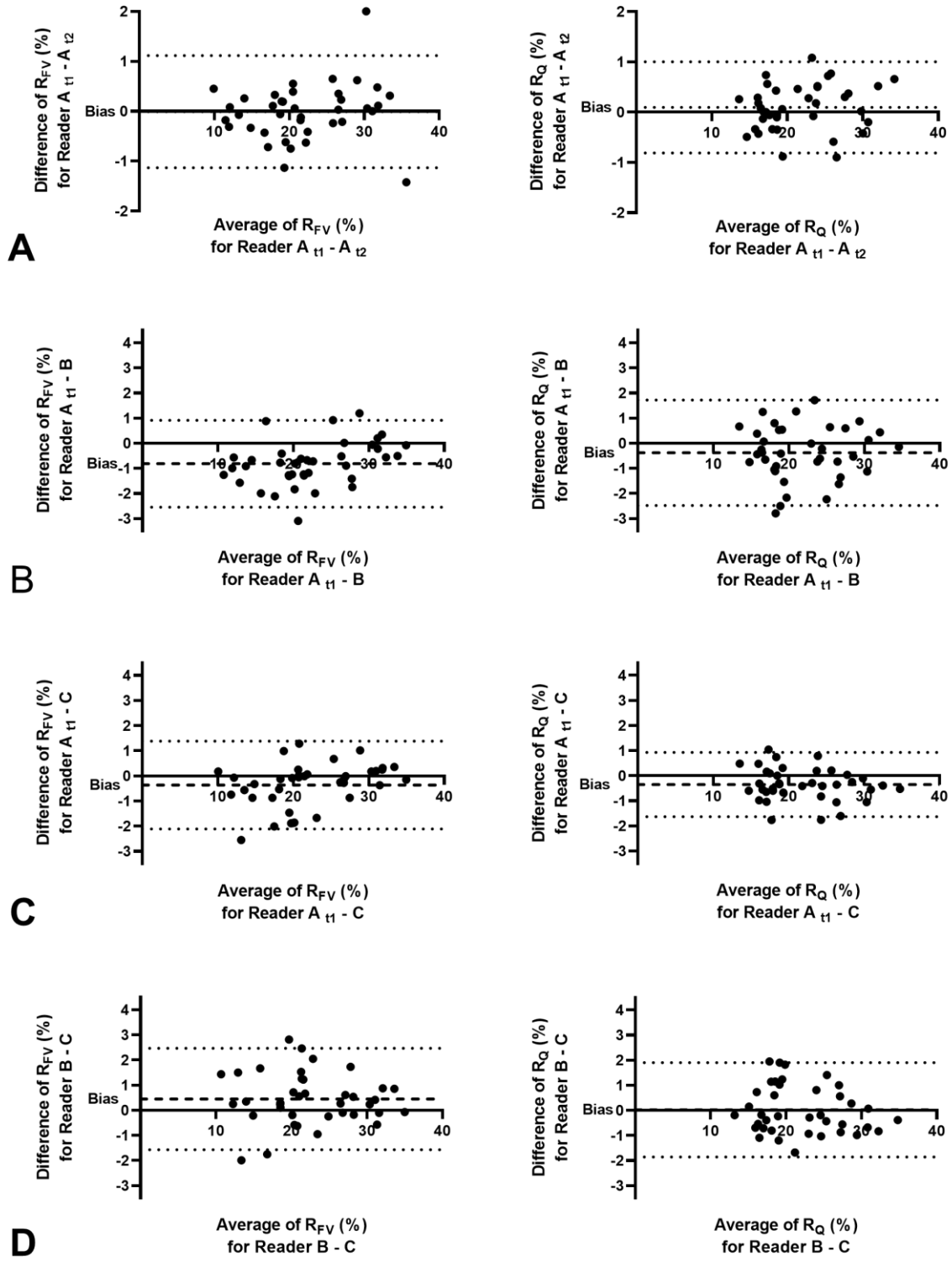


Figure 5. Bland Altman plots for absolute difference between observers. Dashed line represents the bias. Dotted lines represent 95% limits of agreement (LOA). Left column: Ventilation (R_{FV}); Right column: Perfusion (R_Q). Top row: Intra-Observer repeatability A_{t1}- A_{t2}; second row: A_{t1} – B, third row: A_{t1} – C; bottom row: B – C. The biases and LOA amongst human

observers are comparable to the human-ANN one. Abbreviations: R_{FV}, impaired relative fractional ventilation; R_Q, impaired relative perfusion.

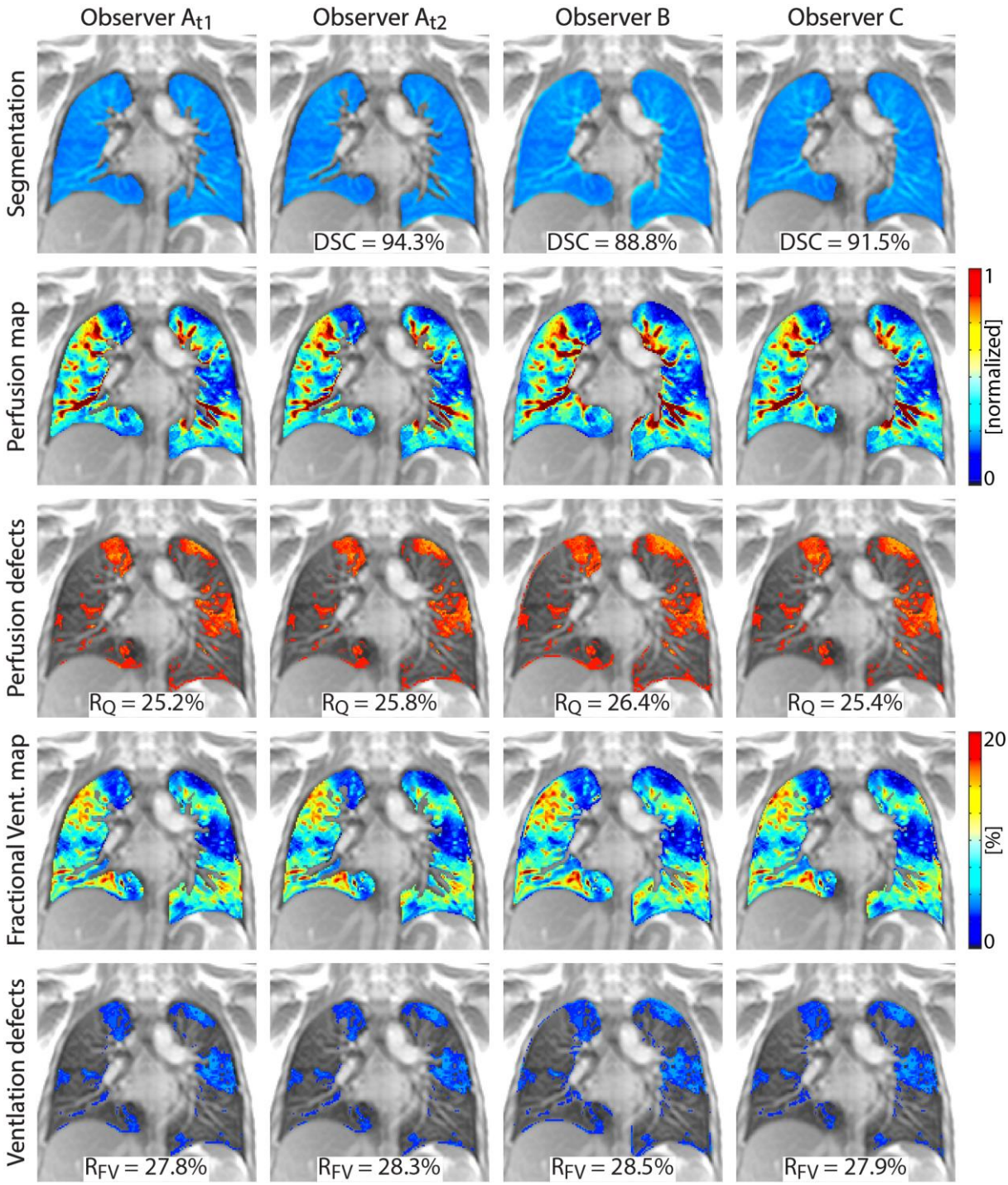
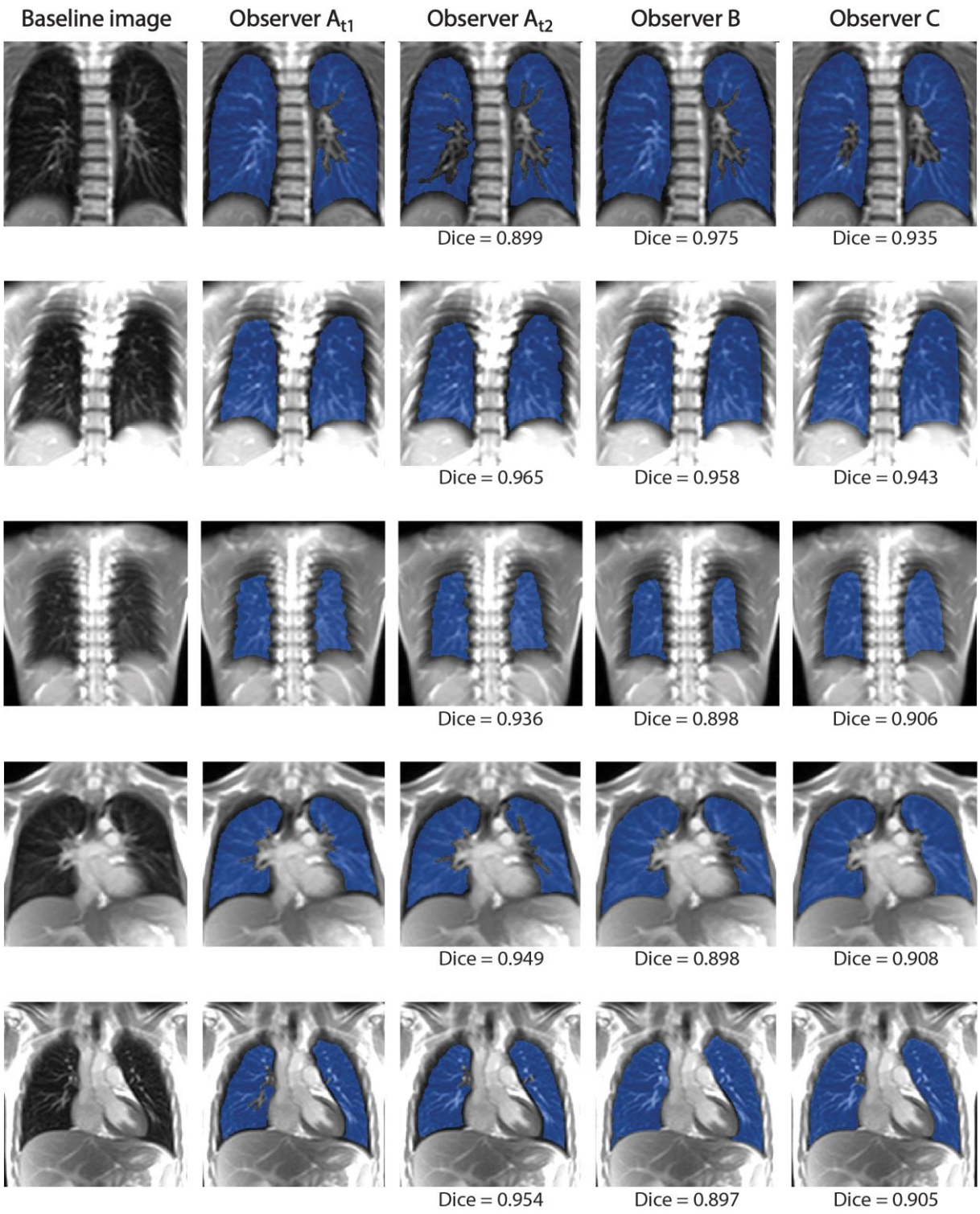


Figure 6. Segmentations of a coronal slice in a 15 year old female with CF (FEV1 z-score= -2.0, LCI = 12.4) performed by the three observers and the resulting perfusion maps, perfusion

defect masks, ventilation maps, and ventilation defect masks. The DSC (calculated with the segmentation A_{t1} as reference), R_Q and R_{FV} are indicated in the figure. The DSC was 87.0% (A_{t2} vs B), 90.3% (A_{t2} vs C), and 92.7% (B vs C). This subject with CF has both ventilation and perfusion defects prominently localized in the apical lung, middle lobe and lingula. To note, the segmentations performed by human observers (A and B) by means of the region-growing algorithm (cf. Methods section) included few pixels on the diaphragm region, while the trained ANN (C) is more precise. Abbreviations: DSC, dice similarity coefficient; R_{FV}, impaired relative fractional ventilation; R_Q, impaired relative perfusion.



Supporting Information Figure S1. Exemplary lung segmentations in healthy and cystic fibrosis subjects. The dice coefficient is calculated using observer A_{t1} as reference.